# LQCD-ext Technical Performance of FY2013 Cluster Deployment

Amitoj Singh

Fermilab

amitoj@fnal.gov

SC LQCD-ext Annual Progress Review
Fermi National Accelerator Laboratory

May 15-16, 2014

# Talk Outline

- Overview of SC LQCD-ext acquisitions
- FY13 conventional cluster deployment and performance
- Questions

# Overview of SC LQCD-ext Acquisitions

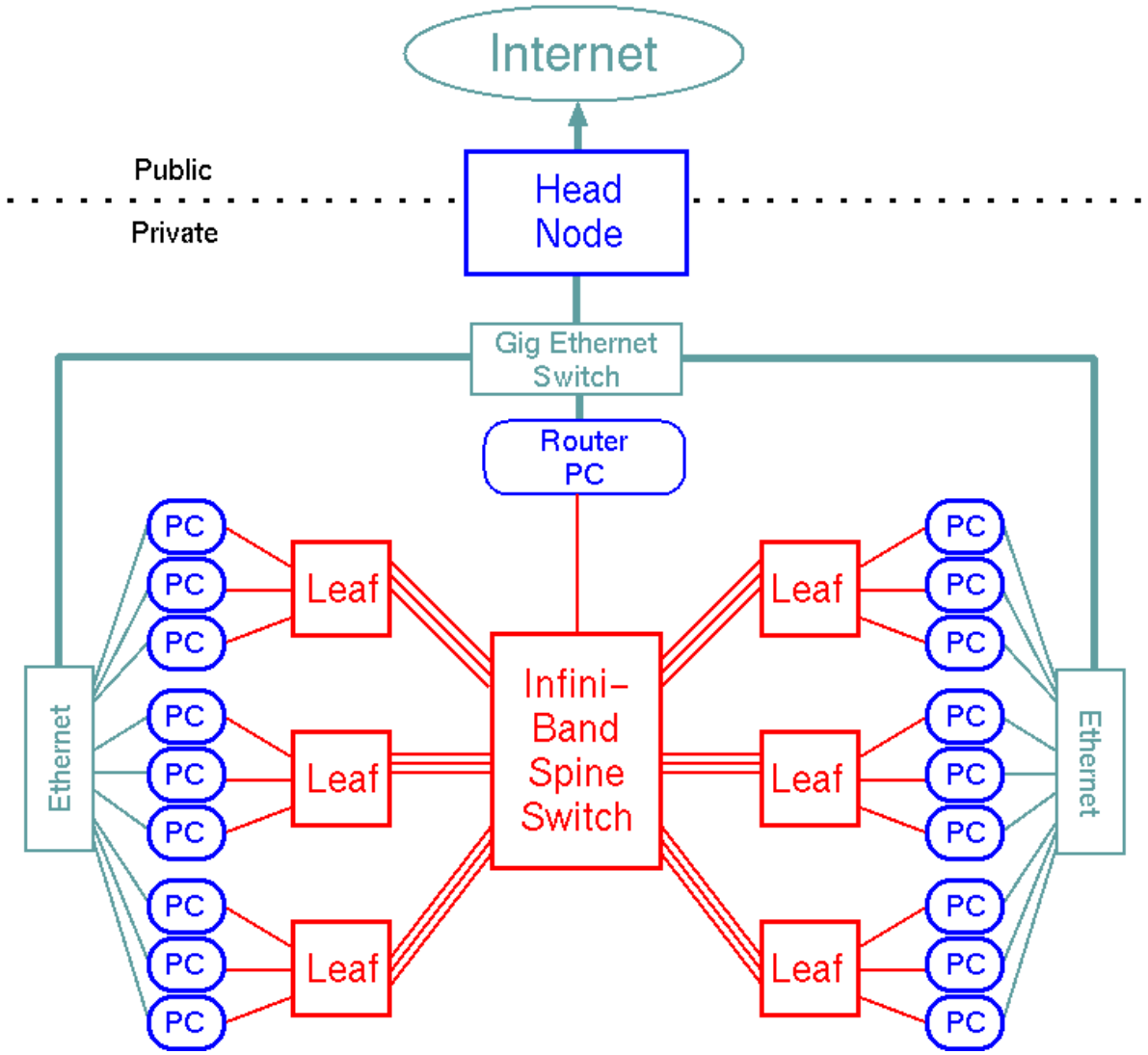
Computational capacity goals by year for SC LQCD-ext:

| | FY2010 | FY2011 | FY2012 | FY2013 | FY2014 |
|---|---|---|---|---|---|
| Baseline computing hardware budget (not including storage) | $1.60M | $1.69M | $1.875M | $2.46M | $2.26M |
| Capacity of new cluster deployments (Tflop/s) Planned/Revised/Achieved | 11 / 12.5 | 12 / 9 / 9 | 24 / 10-15 /12.8 | 44 / 15-22 / 34.6 | 57 / 22-33 |
| Million "Fermi" GPU-Hrs/Yr Planned/Revised/Achieved | 0 | 0 / 1.02 / 1.22 | 0 / 2.9-4.3 / 2.1 | 0 / 4.6-6.9 / 0 | 0 / 7.5-11.2 |

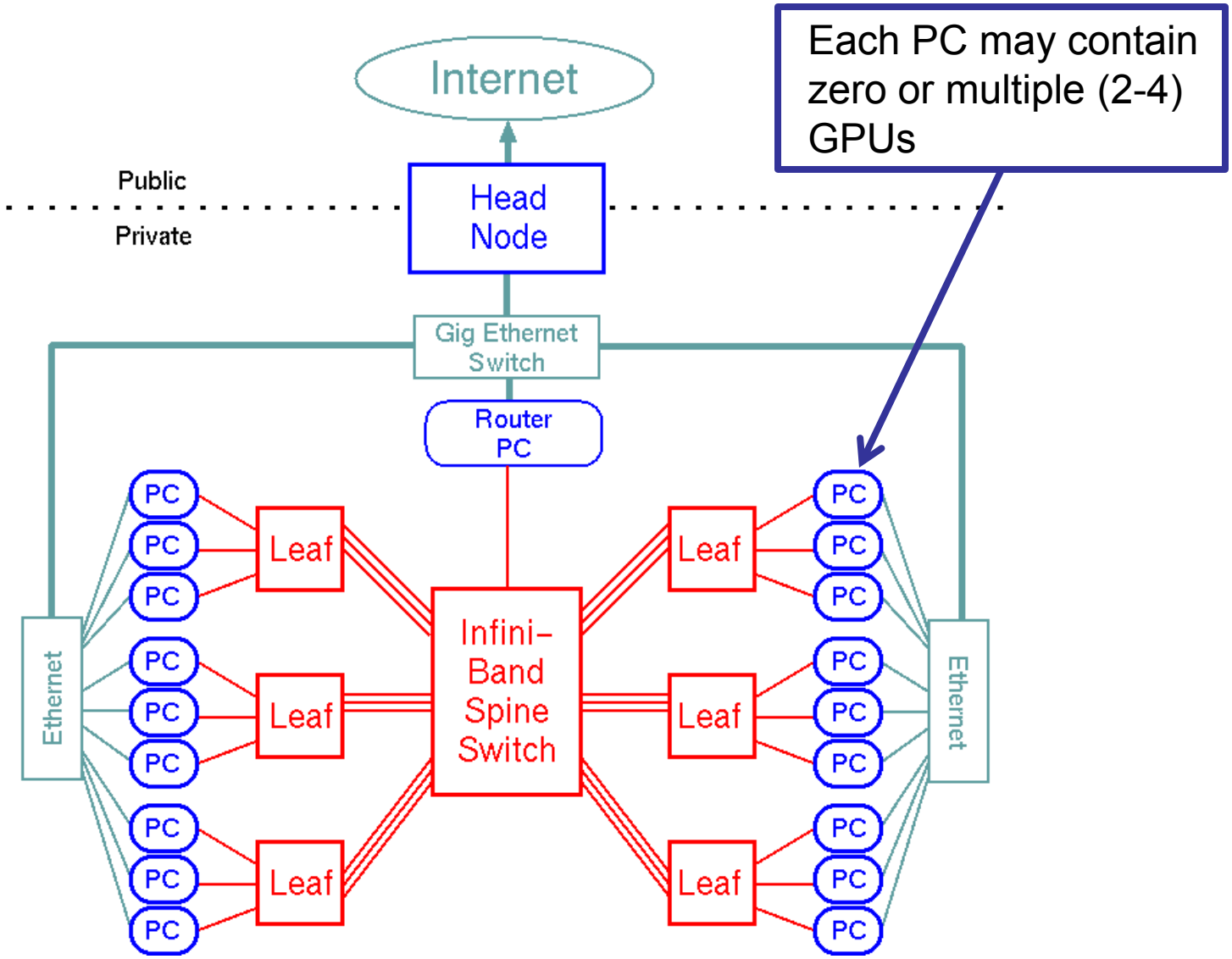
- FY2011 baseline plan for 12 Tflop/s was modified to 9 Tflop/s plus a GPU-accelerated cluster with "Fermi" GPUs (goal 128 GPUs, achieved 156)
- FY2012-FY2014 revised goals reflect 40%-60% ranges in budget allocated to conventional and accelerated clusters. GPU capacity range was extrapolated from the FY2011 purchase using the observed Moore's Law halving time for conventional hardware
- FY2013: project did not deploy GPUs, but a BG/Q half-rack (21.9 TF) and the "Bc" conventional cluster (12.7 TF) that is part of this talk.

# FY13 Conventional Cluster Deployment and Performance

# Typical LQCD Cluster Layout



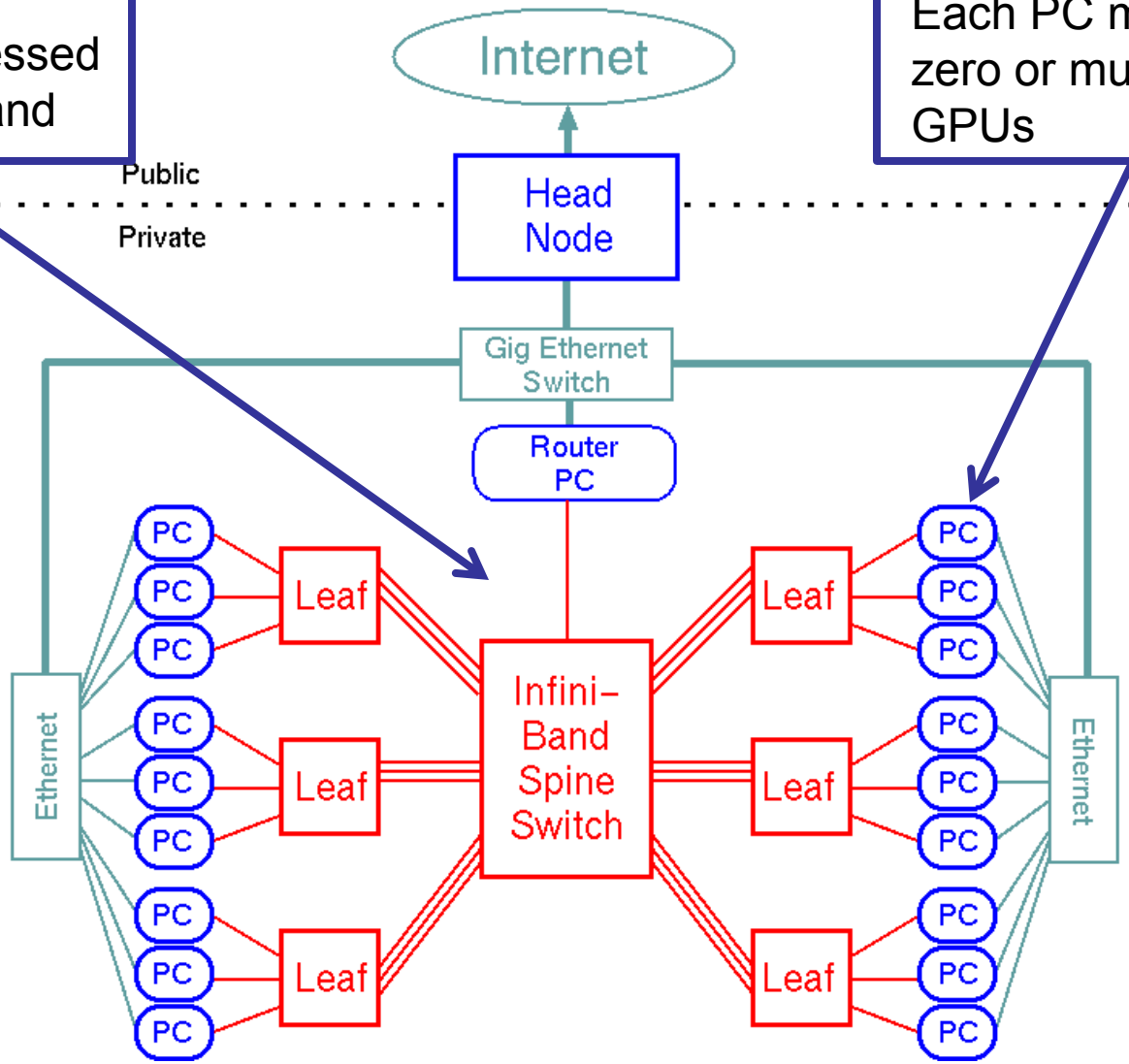
# Typical LQCD Cluster Layout



# Typical LQCD Cluster Layout

Lustre parallel filesystem accessed through Infiniband

Each PC may contain zero or multiple (2-4) GPUs



# Bc Details

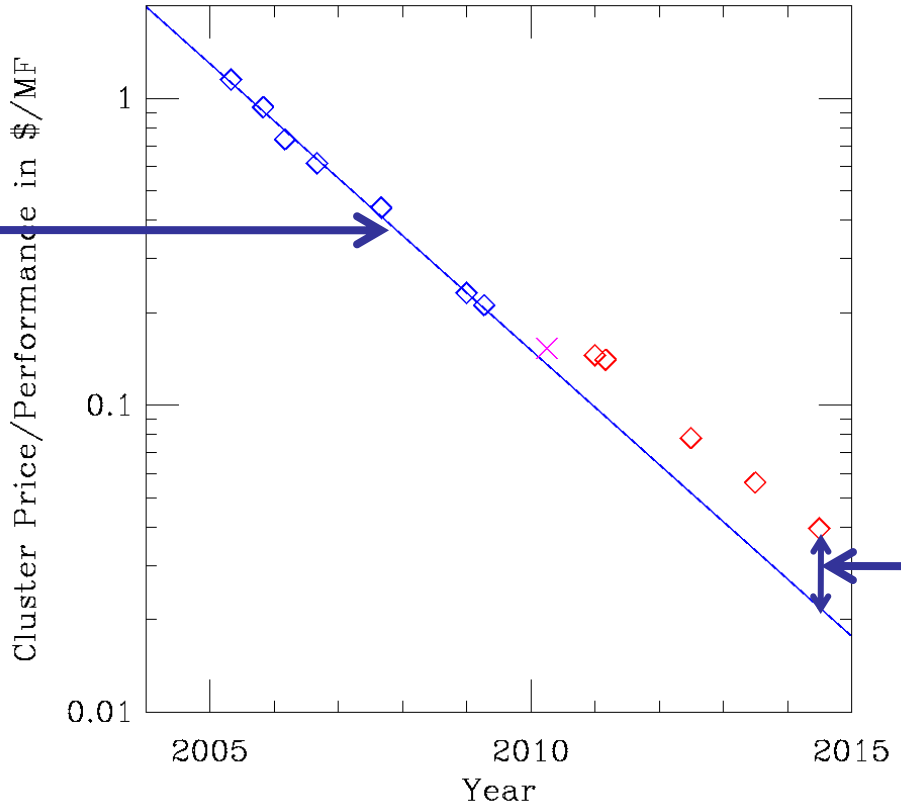
- Award was to best value bid, based on price, LQCD application performance, power efficiency, space efficiency, vendor qualifications and past performance
- Hardware details:
  - Quad-socket eight-core AMD 2.8 GHz “Abu-Dhabi” processors,
  - 64 Gigabytes memory per node,
  - QDR Infiniband with 2:1 oversubscription,
  - 224 worker nodes (7168 cores), plus head nodes and Lustre router nodes,
  - $0.85 M including G&A ($0.75 M for worker nodes + Infiniband)
- Performance
  - Asqtad:DWF  56 Gflop/node (128-process MPI run)
  - 12.7 Tflop/s → $0.056/Mflop

# Bc Details

- Vendors had the option to re-use existing racks, PDUs and contract with on-site DELL Managed Services for the installation and support.
- Housed in a room 300ft  away from existing LQCD clusters.
- Worker nodes connected to Lustre servers via Lustre router nodes using four 10 GigE links.
- Winning bid included Qlogic (Intel-brand) Infiniband. All our current Infiniband was Mellanox based.


# Cost and Performance Basis

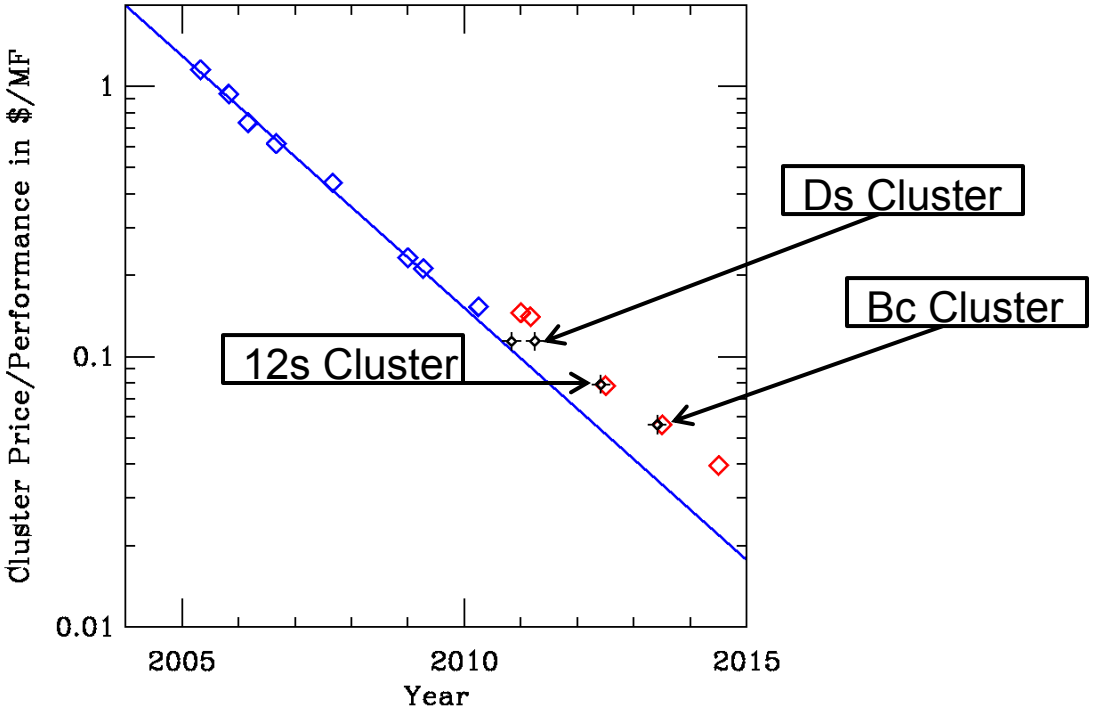
Fit is to the blue diamonds, slope gives halving time of 1.613 years




Contingency

| Year | Deploy Date | Price/Perf. Goal | Price/Perf. Trend | Goal (TF) | Contingency (TF) | Contingency (TF %) |
|---|---|---|---|---|---|---|
| 2010 | 2011.0 | $0.15/MF | $0.098/MF | 11 | 4.4 | 40% |
| 2011 | 2011.2 | $0.14/MF | $0.098/MF | 12 | 4.4 | 36% |
| 2012 | 2012.5 | $0.078/MF | $0.052/MF | 24 | 11.9 | 50% |
| 2013 | 2013.5 | $0.056/MF | $0.034/MF | 44 | 26.8 | 61% |
| 2014 | 2014.5 | $0.040/MF | $0.022/MF | 57 | 42.6 | 75% |

# Cost and Performance Basis

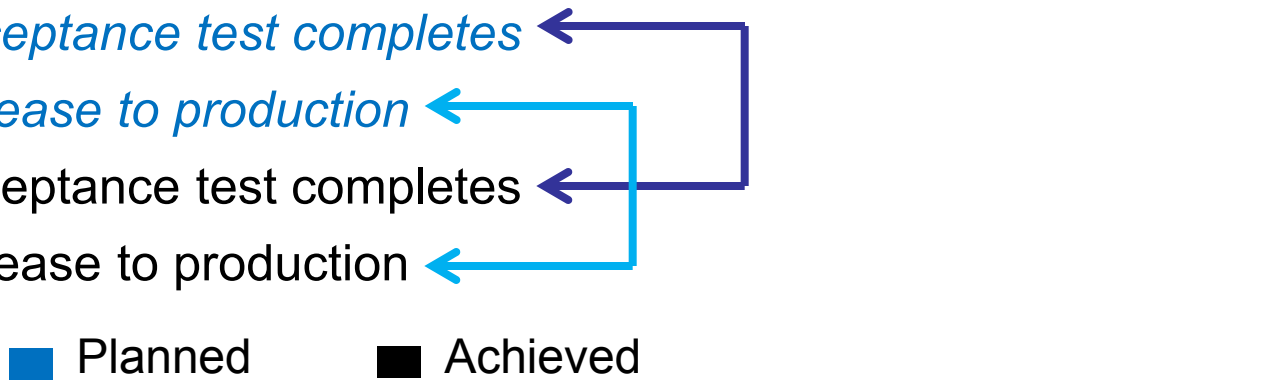



| Cluster | Price per Node | Performance/Node, MF | Price/Performance |
|---|---|---|---|
| Pion #1 | $1910 | 1660 | $1.15/MF |
| Pion #2 | $1554 | 1660 | $0.94/MF |
| 6n | $1785 | 2430 | $0.74/MF |
| Kaon | $2617 | 4260 | $0.61/MF |
| 7n | $3320 | 7550 | $0.44/MF |
| J/Psi #1 | $2274 | 9810 | $0.23/MF |
| J/Psi #2 | $2082 | 9810 | $0.21/MF |
| 10q | $3461 | 22667 | $0.15/MF |
| Ds | $5810 | 50810 | $0.114/MF |
| 12s | $3675 | 50118 | $0.079/MF |
| Bc | $3219 | 56281 | $0.057/MF |

# The FY13 Bc Procurement Timeline

2013

- Jan 11 RFP released to vendors
- Feb 8 Proposal due date
- Mar 1 Purchase Order to vendor
- Apr 24 Switches and rails installed in racks
- *Apr 26 Delivery of all equipment to Fermilab*
- *May 10 Completion of integration. Acceptance test begins*
- May 24 Delivery of all equipment to Fermilab
- June 3 Completion of integration. Acceptance test begins
- *June 10 Acceptance test completes*
- *July 1 Release to production*
- July 1 Acceptance test completes
- July 10 Release to production



# Questions?